Multi-scale Discriminant Saliency with Wavelet-based Hidden Markov Tree Modelling

Anh Cat Le Ngo[a,b], Kenneth Li-Minn Ang[c], Jasmine Kah-Phooi Seng[d],
Guoping Qiu[b], Wee Gin Lim[a]

[a]School of Engineering, The University of Nottigham, Malaysia Campus, 43500 Jalan Broga, Semenyih Selangor, Malaysia
[b]School of Computer Science, The University of Nottingham, Jubilee Campus, Nottingham NG8 1BB United Kingdom
[c]Centre for Communications Engineering Research, Edith Cowan University, 270 Joondalup Dr Joondalup WA 6027, Australia
[d]Department of Computer Science & Networked System, Sunway University, Jalan Universiti Bandar Sunway, 46150 Petaling Jaya, Selangor

Abstract

Bottom-up saliency, an early stage of human visual attention, can be considered as a binary classification problem between centre and surround classes. Discriminant power of features for the classification is measured as mutual information between distributions of image features and corresponding classes . As the estimated discrepancy very much depends on considered scale level, multi-scale structure and discriminant power are integrated by employing discrete wavelet features and Hidden Markov Tree (HMT). With wavelet coefficients and Hidden Markov Tree parameters, quad-tree like label structures are constructed and utilized in maximum a posterior probability (MAP) of hidden class variables at corresponding dyadic sub-squares. A saliency value for each square block at each scale level is computed with discriminant power principle. Finally, across multiple scales is integrated the final saliency map by an information maximization rule. Both standard quantitative tools such as NSS, LCC, AUC and qualitative assessments are used for evaluating the proposed multi-scale discriminant saliency (MDIS) against the well-know information based approach AIM on its released image collection with eye-tracking data. Simulation results are presented and analysed to verify the validity of MDIS as well as point out its limitation for further research direction.

1. Visual Attention - Computational Approach

Visual attention is a psychological phenomenon in which human visual systems are optimized for capturing scenic information. Robustness and efficiency of biological devices, the eyes and their control systems, visual paths in the brain have amazed scientists and engineers for centuries. From Neisser [1] to Marr [2], researchers have put intensive effort in discovering attention principles and engineering artificial systems with equivalent capability. For decades, this research field has been dominated by visual attention principle , proposing an existence of a saliency map for attention guidance. The idea is further promoted in Feature Integration Theory (FIT) [3] which elaborates computational principles of saliency map generation with centre-surround operators and basic image features such as intensity, orientation and colour. Then, Itti et al. [4] implemented and released the first complete computer algorithms of FIT theory [1].

Feature Integration Theory is widely accepted as principles behind visual attention partly due to its utilization of basic image features such as colour, intensity, and orientation. Moreover, this hypothesis is supported by several evidences from psychological experiments. However, it only defines theoretical aspects of saliency maps and visual attention , but does not investigate how such principles would be implemented algorithmically. This lack of implementation details leaves research field open for many later saliency algorithms [4, 5, 6, 7], etc. Saliency might be computed as a linear contrast between features of central and surrounding environments across multiple scales by centre-surround operators. Saliency is also modelled as phase difference in Fourier Transform Domain [8], or its value depends on statistical modelling of the local feature distribution [6]. Though many approaches are mentioned in the long and rich literature of visual saliency, only a few are built on a solid theory or linked to other well-established computational theory. Among the approaches, Neil Bruce's work [9] nicely established a bridge between visual saliency and information theory. It puts a first step for bridging two alien fields; moreover, visual attention for first time could be viewed as information system. Then, information based visual saliency has continuously been investigated and developed in several works [10, 11, 12, 13]. The distinguish points between these works are computational approaches for retrieving information from features. The process attracts much inter-

---

[1]http://ilab.usc.edu/toolkit/

est due to challenges in estimating information of high-dimensional data like 2-D image patches. It usually runs into computational problems which can not be efficiently solved due to the curse of dimensionality; moreover, central and surrounding contexts are usually defined in ad-hoc manners without much theoretical supports. To tackles the problems, Danash Gao et al. has simplified the information extraction step as a binary classification problem with decision theory . Two classes are identified as centre and surround contexts then their discriminant power or mutual information between features and classes are estimated as saliency values for each location. This formulation is named as Discriminant Saliency (DIS) of which underlying principles are carefully elaborated by Gao et al. [14]. Its significant point is estimating information from class distributions rather than from the input features themselves. Therefore, computational load is greatly reduced as only simple class distribution need estimating rather than complex feature distribution.

Spatial features have large influence on saliency values; however, scale-space features do have decisive role in visual saliency computation since center or surround environments are simply processing windows with different sizes. In signal processing, scale-space and spectral space are two sides of a coin; therefore, there is a strong relation between scale-frequency-saliency in visual attention problem. Several researchers [15, 16, 17, 18] outlined that fixated regions have high spatial contrast or showed that high frequency edges allow stronger discrimination between fixated over non-fixated points. In brief, they all come up with one conclusion: increment in predictability at high-frequency features. Although these studies emphasizes a greater visual attraction to high frequencies (edges, ridges, other structures of images), there are other works focusing on medium frequency. Bruce et al. [19] propose that fixation points tend to prefere horizontal and vertical frequency content rather than random position, and these oriented contents have more noticeable difference in medium frequencies. More interestingly, choices of frequency range for biological vision may depend on encountering visual context [20]. For example, luminance contrast explains fixation locations better in natural image category and slightly worse in a category of urban scenes provided that all images are applied low-pass filters as preprocessing steps. Perhaps, that attention system may include different range of frequencies for optimal eye-movements. In other words, diversity in spectral space means necessary utilization of scale-space theory and visual attention. It can be assumed that both high (small scale) and medium frequency (medium scale) constitutes an ecological relevance and balances between information require-

3

ment and available attentional capacity in the early stage of visual attention when observers are not driven by performing any specific tasks.

Though multi-scale nature has been emphasized as the implicit part of human visual attention, it is often ignored in several saliency algorithms. For example, DIS approach [14] considers only one fixed-size window; hence, it may lead to inconsideration of significant attentive features in a scene. Therefore, DIS approach needs constituting under the multiple level framework to form multiscale discriminant saliency (MDIS) approach. This is the main motivation as well as contribution of this paper which are organized as follows. Section 2 reviews principles behind DIS [13] and focuses on its important assumption and limitation. Section 3 provides alternative interpretation of center-surround operation in statistical manner; moreover, it also describes the operation in the light of "large variance" / "small variance" Gaussian distribution as well as "implicit" / "explicit" manifolds concepts. After that, MDIS approach is carefully elaborated in section 4 with several relevant contents such as multiple dyadic windows for binary classification problem in subsection 4.1, multi-scale statistical modelling of wavelet coefficients and learning of parameters in sub-sections 4.2, 4.3, maximum likelihood (MLL) and maximum a posterior probability (MAP) computation of dyadic sub-squares in subsections 4.4, 4.5. Then, all MDIS steps are combined for final saliency map generation in subsection 4.6. Quantitative and qualitative analysis of the proposed method with different modes are discussed in section 5; moreover, comparisons of the proposed MDIS and the well-known information-based saliency method AIM [9] simulation data are presented with several interesting conclusions. Finally, main contributions of this paper as well as further research direction are stated in the conclusion section 6.

2. Visual Attention - Discriminant Saliency

Saliency mechanism plays a key role in perceptual organization, recently several attempts are made to generalize principles for visual saliency. In the decision theoretic point of view, saliency is regarded as power for distinguishing salient and non-salient classes; moreover, it combines classical centre-surround hypothesis with derived optimal saliency architecture. In other word, saliency of each image location is identified by the discriminant power of a feature set with respect to the binary classification problem between center and surround classes. Based on decision theory, this discrim-

inant saliency detector can work with variety of stimulus modalities, including intensity, color, orientation and motion. Moreover, various psychophysics property for both static and motion stimuli are shown to be accurately satisfied quantitatively by DIS saliency maps.

Perceptual systems evolve for optimally producing decisions about the state of surrounding environments with minimum probability of error in a decision-theoretic sense as well as minimum computational effort. In order to achieve these goals, artificial systems needs a mathematical framework for optimization. Mathematically, the problem can be defined as (1) a binary classification of interest stimuli and null hypothesis (salient against non-salient features) and (2) measurement of discriminant power from extracted visual features as saliency at each location in the visual field. The discriminant power is estimated in classification process with respect to two classes of stimuli: stimuli of interest and null stimuli of all uninterested features. Each location of visual field can be classified whether it includes stimuli of interest optimally with lowest expected probability of error. From pure computational standpoint, the binary classification for discriminant features are widely studied and well-defined as tractable problem in the literature. Moreover, the discriminant saliency concept and the decision theory appear in both top-down and bottom-up problems with different specifications of stimuli of interest [5],[13].

The early stages of biological vision are dominated by the ubiquity of "centre-surround" operator; then, bottom-up saliency is commonly defined as how certain the stimuli at each location of central visual field can be determined against other stimuli in its surround. In other words, "centre-surround" hypothesis is a natural binary classification problem which can be solved by well-established decision theory. In this problem, classes can be defined as follows.

- Interest hypothesis: observations within a central neighborhood W_l^1 of visual fields location l.

- Null hypothesis: observations within a surrounding window W_l^0 of the above central region.

At each location, likelihood of either hypothesis depends on the visual stimulus, of a predefined set of features X. The saliency at location l should be measured as discriminating power of features X in W_l^1 against features X in

W_l^0. In other words, discriminant saliency value is proportional to distance between feature distributions of centre and surrounding classes.

In mathematics, feature responses within the windows are drawn from the predefined feature sets X in a process. Since there are many possible combinations and orders of how such responses are assembled, the observations of features can be considered as a random process of dimension d. $X(l) = (X_1(l), \ldots, X_d(l))$. This random process is drawn conditionally on the states of hidden variable $Y(l)$, which is either centre or surround state. Feature vector $\mathbf{x}(\mathbf{j})$ such that $j \in W_l^c, c \in \{0, 1\}\}$ are drawn from classes c according to conditional densities $P_{X(l)|Y(l)}(x|c)$ where $Y(l) = 0$ for surround or $Y(l) = 1$ for centre. The saliency S at location l, $S(l)$, is equal to the discriminant power of X for classifying observed feature vectors, which can be quantified by the mutual information between feature, X and class label, Y.

$$S(l) = I_l(X; Y) = \sum_c \int p_{X,Y}(x, c) log \frac{p_{X,Y}(x, c)}{p_X(x) p_y(c)} dx$$

. Though binary classification and decision theory makes discriminant saliency computationally feasible, it is only for low-dimensional data. Computer vision and visual attention need to deal with high-dimension input images especially when it involves statistics and information theory. As mentioned previously, observations of feature responses, $X(l)$, are considered as a random process in d-dimensional space. Mutual information estimation of high-dimension data encounters serious obstacles due to the curse of dimensionality as well as computational efficiency. As these problems persist, saliency algorithms would never be biologically plausible and computationally feasible. Therefore, discriminant saliency algorithms have to be approximated by taking into account statistical characteristics of natural images as well as mathematical simplification. Dashan Gao and Nuno Vasconcelos have proposed a feasible algorithm for mutual estimation, mathematically formulated

as follows.

$$
\begin{aligned}
I_l(X;Y) &= H(Y) - H(Y|X) \\
&= E_X(H(Y) + E_{Y|X}[logP_{Y|X}(c|x)]) \\
&= E_X \left[H(Y) + \sum_{c=0}^{1} P_{Y|X}(c|x)logP_{Y|X}(c|x) \right] \\
&= \frac{1}{|W_l|} \sum_{j \in W_l} \left[H(Y) + \sum_{c=0}^{1} P_{Y|X}(c|x_j)log\,P_{Y|X}(c|x_j) \right]
\end{aligned}
\tag{1}
$$

where $H(Y) = -\sum_{c=0}^{1} P_Y(c)logP_Y(c)$ is the entropy of classes Y and $H(Y|X) = -E_{Y|X}\left[logP_{Y|X}(c|x)\right]$ is the conditional entropy of Y given X. Given a location l, there are corresponding centres W_l^1 and surround W_l^0 windows along with a set of associated feature responses $x(j), j \in W_l = W_l^0 \cup W_l^1$. The mutual information can be estimated by replacing expectations with means of all samples inside the join windows W_l. The conditional entropy $H(Y|X)$ can be computed by analytically deriving MAP $P(Y|X)$ given deployment of Generalized Gaussian Distribution (GGD) for transformed features in a binary classification problem. Let's name Gao's proposal for discriminant saliency computation as DIS; more details about DIS can be found in their publications [14][13][21][22].

While DIS successfully defines discriminant saliency in decision theoretic sense, its implementation has certain limits. Feature responses are randomly sampled in a single fixed-size window; therefore, it is obviously biased toward objects with distinctive features fitted in that window size. As previous findings have confirmed involvement of multi-scale factors in visual attention, DIS needs extending from a fixed-scale process to a multi-scale process. In theory, DIS can be carried out with different size of windows, and this approach certainly produces image responses and saliency values at multiple scales. However, such approach is not recommended for both computational and biologically mechanisms since it causes high redundancy in saliency values across multiple scales. In order to solve the multi-scale problem systematically, DIS should be integrated with multiple scale processing techniques such as wavelet transforms.

3. Visual Attention - Centre and Surround Context

Discriminant Saliency (DIS) studies differences between two pixel distributions sampled from assigned windows called centre and surround windows. It intuitively fits into the concept of centre and surround operators, and the idea would work well if linear difference such as image contrast needs figuring out. However, deployment of mutual information, a high-order comparison between two distributions would require enormous amount of sampled pixels which enclosed image patches certainly can not afford. Therefore, centre-surround operators have to be considered in the informatics and statistics' point of view.

Modelling 2-D data such as images is a complicated task due to infinite number of possible joint distributions. However, in the visual attention study, natural images are the only image class which are analysed. Then, it simplifies the task significantly due to numerous works on statistics of natural images in the literature [23, 24, 25, 26, 27, 28]. From several studies about statistical aspects of natural images, we notice importance of explicit and implicit manifolds by Shi and Zhu [29]. They have argued image patches as fundamental elements for objects modelling and recognition in natural images. Moreover, these patches plays a key role in forming the whole ensemble of natural image patches which are in turn classified into two types of subspaces: explicit and implicit manifolds. The explicit manifolds mainly contain simple and regular image primitives such as edges, bards, corners, and junctions while implicit manifolds are dominated with complex and stochastic image patches such as textures and clutters. Image scaling realizes a connection between two types of manifolds; in other words, a specific point on images can be classified as either edge or texture according to the patch size. Furthermore, image scales not only changes nature of image patches from implicit to explicit manifolds but as well affect their entropy since two subspaces live separately in low and high entropy regimes. Studying this transition over scale shows the peak of manifold complexity in the middle entropy regime [29]. In other words, there exists an intermediate scale which collects the most information of patches if information is measured by entropy value.

The transition from implicit to explicit manifolds or vice verse over multiple scales in natural images have strong correlation with centre-surround operators of visual attention theory. Centre or surround windows considerably correspond to smaller patches and bigger patches at the same location; in other words, they are patches at two consecutive scales. When two or
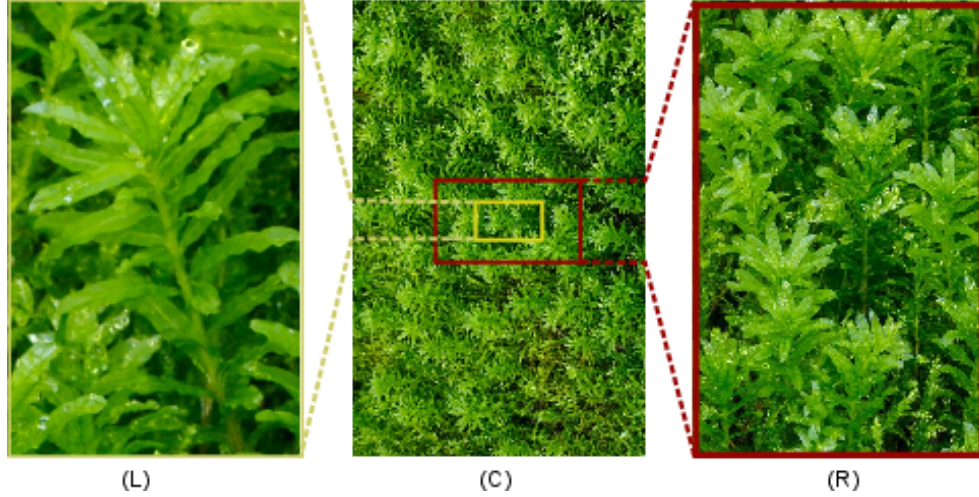
Figure 1: (L) : central contex (explicit manifold). (C): original image. (R): surrounding context (implicit manifold)

more pairs of centre and surround patches are employed, multi-scale centre-surround operators are formed with blocks at different scales. As image scaling plays a key role in classification of patches into either implicit or explicit manifold, centre and surround patches have different likelihood towards the manifolds. Therefore, DIS can be formulated as binary classification between two different types of manifolds with statistical constrains instead of two distinguishing patches with geometrical constrains. It certainly benefits estimation of mutual information because DIS with statistical constrains can be learn from patches across images, the global context while the original DIS with geometrical constraint restrains the accuracy of information estimation due to its small number of pixels in local contexts.

Deployment of DIS approach in deciding classes of image patches according to their statistical characteristics need mathematical models of information estimation for realization. Shi and Zhu proposes manifold pursuits along with the introduction of implicit and explicit manifolds. The pursuit utilizes iterative approaches EM algorithms to produce approximate statistical models of image patches ($\mathbf{I}$) over the global context Ω. In their approaches, training samples are image pixels, spatial features, which is again difficult for multi-scale extension. Alternatively, we try to model statistical characteristics of wavelet features with Gaussian Mixture Models (GMM), a good statis-

9

tical approximation of wavelet coefficients distributions from natural images as well as their usage of large-variance and small-variance Gaussian distributions. Again, these large/small variance or implicit/explicit distributions are just two aspects of one phenomenon assumed that Gaussian distributions are suitable for modelling features of natural images. For instances, large variance Gaussian distribution means large uncertainty or high-entropy regime where implicit manifold live on; meanwhile, small variance Gaussian distribution has small uncertainty or low entropy regime where explicit manifolds exists. Therefore, statistical constrains for DIS in the global context can utilize large/small-variance Gaussian distributions instead of implicit/explicit manifolds concepts due to its convenience in mathematical models. In this paper, GMM concepts are seen through for modelling wavelet coefficients, spatial-scale features.

4. Multiscale Discriminant Saliency

Expansion from a fixed window-size to multi-scale processing is commonly desired in development of computer vision algorithms. In long literature of computer vision research fields, there are many multi-scale processing models applicable for a so called Multiscale Discriminant Saliency (MDIS). Any selection has to adapt binary classification in multi-scale stages. Put differently, it requires efficient classification of an image datum into a class at a particular scale with prior knowledge from other scales. With respect to these requirements, a multi-scale image segmentation framework should be a great starting point for MDIS since DIS can be considered as simplified binary image segmentation with only two classes. However, the binary classification is only an intermediate step to measure discriminant power of center-surrounding features, and accuracy of segmentation results does not really matter in this case.

Typical algorithms employ a rigid classification window in a vague hope that all included pixels belong to the same class. Obviously, DIS has similar problems of choosing suitable window sizes as well. Clearly, the size is crucial to balance between its reliability and accuracy. A large window usually provides rich statistical information and enhance reliability of the algorithm. However, it also risks including heterogeneous elements in the window and eventually loses segmentation accuracy. Therefore, appropriate window sizes are equivalently vital in avoidance of local maxima in discriminant power. If window sizes are too large or too small, MDIS risks losing useful discrimina-

tive features or being too susceptible to noise. In brief, sampling rates, and consequently number of data in a window, directly affect on performance of binary classification / segmentation and eventually computation of discriminating powers.

4.1. Multiple Classification Windows

Multi-scale segmentation employs multi-scale classification windows on images, then combines responses across scales. MDIS can adapt similar approach to model classification image features into either centre or surrounding classes. Though window sizes are preferred to be chosen arbitrarily, dyadic squares (or blocks) are implemented in MDIS because of its compactness and efficiency. Let's assume an initial square image s with 2^Jx2^J of $n := 2^{2J}$ pixels, the dyadic square structures can be generated by recursively dividing x into four square sub-images equally, left-side of the figure 2. As a result, it has the popular quad-tree structure, commonly employed in computer vision and image processing problems. In this tree structure, each node is related to a direct above parent node while it plays a role of parental nodes itself for four direct below nodes 2. Each quad-tree node is equivalent with a dyadic square, and is denoted as a tree-node in scale j by d_i^j whereof i is a spatial index of a dyadic square node. Given a random field image X, the dyadic squares are also random fields which are formulated as D_i^j mathematically. In following sections, we sometime use D_i (dropping scale factor j) as general randomly-generated dyadic square regardless of scales. Using these dyadic squares as classification windows, we can classify each node d_i as either centre or surround by estimating its maximum a posterior probability (MAP). Then, mutual information between features and corresponding labels can be computed by averaging MAP across all classes. The mutual information is similar to the core concept of discriminant power of central features against surrounding ones at each location [5]. However, deployment of quad-tree structures makes the estimation of mutual information possible for many scales. In order to compute the discriminant power, multiple PDF need to be learned through wavelet-based statistical models.

4.2. Multi-scale Statistical Model

Hidden Markov Model (HMM) captures the main statistical features of wavelet transforms of real-world images. In one hand, parameters originally have to be learn for each data point in the raw model of HMT; however, such training make it unwieldy for handling images with large amount of data. In

11

Figure 2: Quad-Tree and Dyadic Wavelet Structures

the other hand, arbitrarily specified parameters would help to avoid a time-consuming learning stage which is even impossible in some cases, but would risk of over-fitting the model. The above issues may render HMT inappropriate for applications with rapid processing requirement but without sufficient priori information. Therefore, several derivatives of HMT algorithms are studied in this section in order to realize HMT more feasible manner as well as find out their advantages and limitations.

Marginal distributions of wavelet coefficients w_i come directly from sparseness of wavelet transform in modelling real-word images: a minority of large coefficients and a majority of small coefficients. That distribution is efficiently captured by Gaussian Mixture Modelling (GMM) with wavelet coefficients w_i and observed hidden state variables or class labels $S_i \in S, L$. State

value S_i decides which mixture generates coefficients w_i.

$$g(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi}\sigma} exp - \frac{(x-\mu)^2}{2 * \sigma^2}$$

Lets denotes the Gaussian PDF of small-variance state $S_i = S$ is as follows.

$$f(w_i | S_i = S) = g(w_i; 0, \sigma_{S;i}^2)$$

While state $S_i = L$ has zero-mean, large-variance Gaussian

$$f(w_i | S_i = L) = g(w_i; 0, \sigma_{L;i}^2)$$

where $\sigma_L^2 \geq \sigma_S^2$. By those mixture models, we can write the marginal PDF $f(w_i)$ as a convex combination of the conditional densities.

$$f(w_i) = p_i^S g(w_i; 0, \sigma_{S;i}^2) + p_i^L g(w_i; 0, \sigma_{L;i}^2)$$

where $p_i^S + p_i^L = 1$ since $p_{S_i} = [p_i^S p_i^L]$ are mass probability of states. In statistical interpretation, p_{S_i} is how likely wavelet coefficients w_i is small or large.

HMT captures inter-scale dependencies of probabilistic tree connecting hidden state variables of a wavelet coefficient and its four children. Therefore, dependency graphs have similar quad-tree topology as wavelet decomposition, and it includes state-to-state links between parent and child coefficients, mathematically modelled by persistency probabilities and novelty probabilities.

$$A_i = \begin{bmatrix} p_i^{S \to S} & p_i^{S \to L} \\ p_i^{L \to S} & p_i^{L \to L} \end{bmatrix}$$

where $p_i^{S \to S} + p_i^{S \to L} = 1$ and $p_i^{L \to S} + p_i^{L \to L} = 1$. Persistency probabilities are lying on the main diagonal axis of the above array $p_i^{S \to S}, p_i^{L \to L}$, since they represent how likely states are kept in parent and child links. On the sub diagonal axis are novelty of probabilities, which causes different states between parent and child nodes.

In summary, trained HMT models can be specified in terms of (i) GSM mixture variances $\sigma_{S;i}^2$ and $\sigma_{L;i}^2$ (ii) the state transition matrix A_i and (iii) probability mass function p_1 at the coarsest level. Grouping these parameters

into a model $\mathcal{M}_c$, we can define trained HMT model as follows

$$\mathcal{M}_c = \{p_1, A_2, \ldots, A_J; \sigma_{S_i;k,j,b}\} \qquad (2)$$

$$\text{where}$$

$$j = 1, \ldots, J$$
$$S_i = \{L, S\}$$
$$k \in \mathbb{Z}^2$$
$$\forall b \in \mathcal{B}$$

where k is wavelet coefficients index at scale j while b represents wavelet sub-bands.

$$\mathcal{B} = \{HL, LH, HH\}$$

The parametric model $\mathcal{M}_c$ can be used to learn from the joint PDF $f(\mathbf{w}|\mathcal{M}_\mathbf{c})$ of the wavelet coefficients in each of three sub-bands [30]. Generally, each node of wavelet coefficients has its own model with different parameters. However, it overcomplicates the model with too many parameters; for examples, n wavelet coefficients are required to be fitted on 4n parameter models which is an impossible task. Therefore, to reduce the complexity, HMT is assumed to use same parameters for every nodes at the same scale of wavelet transforms regardless of spatial k and oriental b indexes.

$$\sigma^2_{S;b,j,k} = \sigma^2_{S;j}$$
$$\sigma^2_{L;b,j,k} = \sigma^2_{L;j}$$
$$Ab, j, k = A_j$$
$$k \in \mathbb{Z}^2 \quad , \quad \forall b \in \mathcal{B}$$

The assumption is called *tying within scale* [30] and it prevents over-complex and infeasible HMT model in exchange of less general model, which is mathematically formulated as follows.

$$\mathcal{M}_c = \{p_1, A_2, \ldots, A_J; \sigma_{S_i;j}, (j = 1, \ldots, J, S_i = L, S)\} \qquad (3)$$

Lets name the $\mathcal{M}_c$ in equation 3 as trained Hidden Markov Model (THMT). While the *typing* trick has significantly simplified the learning process for HMT parameters. In fact, the model only needs training on an input image. However, it possible to further simplify the approach if image class are known. Using many images with similar contexts, we can train HMT offline for meta-parameters , which are later fixed in an HMT model. It yields a general HMT

model for that class of images with each member in the class being treated statistically equivalent [31]. Fortunately, Romberg .et .al has studied similar models for the class of natural images and publish their universal parameters obtained by (jointly) fitting lines to the HMT parameters of four images as follows.

$$
\mathcal{M}_c = \left\{
\begin{array}{rcc}
\alpha_S & = & 3.1 \\
C_{\alpha_S} & = & 2^1 1 \\
\alpha_L & = & 2.25 \\
C_{\alpha_L} & = & 2^1 1 \\
\lambda_S & = & 1 \\
C_{SS} & = & 2^{2.3} \\
\lambda_L & = & 0.4 \\
C_{LL} & = & 2^0.5 \\
p_{j0}^L & = & 0.5
\end{array}
\right\}
$$

The variance and persistence decays are measured by fitting a line to the log of the variance versus scale for each state [31], and it is only started at scale $j = 4$ with transition state probability at $j = 5$. This choice of scale ensures enough data for an accurate estimate of decays; moreover, Romberg .et .al [31] finds out that the decays are very similar for many of the natural images. Therefore, it is reasonable to fix a HMT model with meta-parameters which are learnt from natural images. This approach of modelling HMT is called Universal HMT (UHMT) in this paper. Though accuracy of this UHMT model is clearly lost by treating all different images statistically equivalent, the assumption can totally eliminate the need for training and save tremendously computational workload, and make real-time HMT possible. Experiments with UHMT mode are mentioned in the section 5 where both accuracy and efficiency of UHMT is evaluated.

As UHMT approach eliminates training stages of THMT to decrease computational requirement for real-time applications, sometimes it is necessary to have better HMT in modelling image. Though *tying* THMT assumes the same parameters for wavelet sub-bands or coefficients at different orientations, its underlying learning treats each brand independently. However, experiments by Simoncelli and Portilla [32] demonstrate importance of cross-correlation between wavelet sub-bands with different orientations at the same scale for modelling texture image. Moreover, textural features are pretty common in natural images as well; therefore, capturing this dependency would improve accuracy of HMT models. Since THMT treats coefficients of sub-bands independently, it obviously ignores the cross-orientation

15

correlation. To enhance the capacity of THMT, Do and Vetterli [27] proposes to group coefficients at the same location and scale into a vector and carry out HMT modelling in a multidimensional manner. In this paper, we call such approach as Vectorized HMT (VHMT). Let's denote a vector after the grouping of wavelet coefficients at location k, scale j in three different orientations vertical (V), horizontal (H), and diagonal (D) as follows.

$$\mathbf{w}_{j,k} = (w_{j,k}^H w_{j,k}^V w_{j,k}^D)^T$$

As multidimensional groups of wavelet coefficients $\mathbf{w}_{j,k}$ are concerned, their distribution need formulating as zero-mean multivariate Gaussian density with covariance matrix C as follows

$$g(\mathbf{w}; C) = \frac{1}{\sqrt{(2\pi)^n |det(C)|}} \exp(-\mathbf{w}^T C^{-1} \mathbf{w})$$

where n is the number of dimensions, or orientations $n = 3$ in this case. Except from a multivariate probability density, VHMT has similar approach as other HMTs do. Its marginal distribution is formulated as Gaussian Mixture Models, i.e.

$$f_j(\mathbf{w}) = p_j^L g(\mathbf{w}; C_j^L) + p_j^S g(\mathbf{w}; C_j^L)$$

Moreover, its statistical inter-scale dependency is modelled through the parent-child relationship with a quad-tree structure linking a parent with its four children at the next level in the same location. A small difference here is that only one tree is utilized instead of three trees since wavelet coefficients have been grouped and modelled simultaneously. Hence, an image can be modelled by VHMT with a set of parameters.

$$\theta = \{\mathbf{p_1}, A_2, \ldots, A_J; C_j^{S_i}, (j = 1, \ldots, J, S_i = L, S)\}$$

As only one quad-tree is built for modelling a wavelet coefficients vector , the hidden state are as well "tied up". It means the same hidden state is assigned regardless of orientations; in other words, VHMT is orientation-invariant. VHMT captures dependencies across orientations via the covariance matrix C of the multivariate Gaussian density. Diagonal elements of the matrix are variances of each orientation meanwhile non-diagonal elements represent covariance of wavelet coefficients across sub-bands. It justifies VHMT model for textural features since their wavelet coefficients have high possibility of being significant at all orientations in edge regions whereas they are likely small at any directions in smooth regions.

All three derivatives (U/T/V)HMT have distinguishing approaches to a common goal: capturing statistical characteristics of natural images. Each approach has aimed to emphasize different aspects; for example, THMT models general parent-child dependency, UHMT eliminates the need for training by utilization of meta-parameters for a certain image category, while VHMT emphasizes on directionally invariant features. Implementation of those derivatives has both positive and negative effects, which are further elaborated in the section 5.

4.3. Multi-scale Statistical Learning

The complete joint pixel PDF is typically overcomplicated and difficult to model due to their high-dimensional nature. Unavailability of simple distribution model in practice motivates statistical modelling of transform-domain which is often less complex and easier to be estimated. Obviously, joint pixel PDF could be well approximated as marginal PDF of transformed coefficients. Since wavelet transform well-characterizes semantic singularity of natural images, it provides a suitable transform-domain for modelling statistical property of singularity-rich images.

Natural images are full of edges, ridges and other highly structural features as well as textures; therefore, wavelet transforms as multi-scale edge detectors well represent such singularity rich contents at multiple scales and three different directions. Noted that, only normal discrete wavelet transform (DWT) is considered in this study for the sake of simplicity though concepts can be adapted into other wavelet transforms as well. Henceforth, whenever a wavelet transform is mentioned, it refers to DWT instead of stating otherwise. As multi-scale edge detectors, responses of wavelet transform overlying a singularity are large coefficients while DWT of a smooth region yields small coefficients. Simple hard-threshold of wavelet coefficients leads to binary differentiation of singularity against non-singularity features. Moreover, statistical model of images can be significant simplified under "restructured" multi-scale singularity representation.

Quad-tree structure of dyadic squares in pixel domain can be mirrored in wavelet decomposition, the right-hand side of figure 2 since four wavelet coefficients at a given scale nest inside one at the next coarser level. For example, Haar wavelet coefficients at each quad-tree nodes are generated by Harr wavelet transform of the corresponding dyadic image square, which is clearly illustrated by projection from a combination of LL_2, HL_2, LH_2, and HH_2 to LL_3. The combination of singularity detector and multi-scale

quad-tree structure implies the singularity property at each spatial location persists through scales along the branches of the quad-tree in the transform domain.

The singularity characterization along scales makes the wavelet domain well-suited for modelling natural images. In fact, statistical models of wavelet coefficients have quite comprehensive literature; however, we only concentrate on the Hidden Markov Tree of Crouse, Nowak and Baraniuk [30]. In consideration of both marginal and joint statistics of wavelet coefficient, the HMT model introduces a hidden state variable of either "large" or "small" to each quad-tree node at a particular scales. Then, the marginal density of wavelet coefficients is modelled as a two-states Gaussian mixture in which a "large" or "small" refers to characteristics of Gaussian distribution's variance values. The mixtures closely match marginal statistics of natural images [33],[34], [35]. With the HMT, persistence of large or small coefficients are captured across scales using Markov-1 chain. It models dependencies between hidden states across scale in a tree structure, parallel to that of wavelet coefficients and dyadic squares. With parameters of Gaussian Mixture Model (GMM) and Markov State Transition in vector $\mathcal{M}$, the HMT model is able to approximate the overall joint pdf of the wavelet coefficients $\mathbf{W}$ by a high-dimensional but highly structured Gaussian mixture models $f(\mathbf{w}|\mathcal{M})$.

Highly structural nature of wavelet coefficients allows efficient implementation of HMT-based processing. The parameters of a HMT model $\mathcal{M}$ can be learned through the iterative expectation and maximization (EM) algorithm with cost $\mathcal{O}(n)$ per iteration [30] in (T/V)HMT or predefined for a particular image category [31] in UHMT. After the parameters $\mathcal{M}$ are estimated by the EM iteration, we need to compute statistical characteristics of wavelet coefficients given the DWT coefficient $\tilde{\mathbf{w}}$ of an image $\tilde{\mathbf{x}}$ and a set of HMT parameters $\mathcal{M}$. It is a realization of the HMT model in which computation of the likelihood $f(\tilde{\omega}|\mathcal{M})$ requires only a simple $\mathcal{O}(n)$ up-sweep through the HMT tree from leaves to root [30].

Conveniently, wavelet-based HMT models use similar structures as wavelet decomposition and dyadic squares do. Therefore, statistical behaviours of each square block $\tilde{\mathbf{d}}_{\mathbf{i}}$ can be approximately computed by a HMT branch rooted at a node i. As mentioned earlier, maximum likelihood of the sub-tree $\mathcal{T}_i$ is computed simply by up-sweeping from corresponding leave-nodes at scale $S = J$ to the root node at scale $S = j$. If the "up-sweeping" operation is done from leave nodes $S = J$ to root node at scale $S = 0$, we could find out likelihood probability of the whole image [30]. The estimation

18

at intermediate scales j gives the statistical behaviour of the corresponding dyadic sub-square ,$f(\tilde{\mathbf{d}}_i|\mathcal{M})$, under the HMT model.

The above model opens a prospect of a simple multi-scale image classification algorithm. Supposed center and surround classes are denoted as $c \in \{1,0\}$, we have specified or trained an HMT trees for each class with parameters $\mathcal{M}_c$. Then the above likelihood calculation is deployed on each node of the HMT quad-tree given the wavelet transform $\tilde{\omega}$ of an image $\tilde{\mathbf{x}}$. For each node of the tree, HMT yields the likelihood $f(\mathbf{d_i}|\tilde{\mathcal{M}}_c),c \in \{1,0\}$ for each dyadic block d_i. With the multi-scale likelihoods at hand, we can easily choose the most suitable class c for a dyadic sub-square $\tilde{\mathbf{d_i}}$ as follows.

$$\hat{c}_i^{ML} := argmax_{c\in\{1,0\}}f(\tilde{\mathbf{d_i}}|\mathcal{M}_c)$$

The most likely label $\hat{c}_i^{ML}$ for each dyadic sub-square $\tilde{\mathbf{d_i}}$ can be found by a simple comparison between estimation of available classes. Moreover, the approach is also capable of handling large number of input data such as images because of linear computational cost, $\mathcal{O}(n)$ operations for an entire n-pixel image.

4.4. Multiscale Likelihood Computation

For a given set of HMT model parameters $\mathcal{M}$, it is straight forward to compute the likelihood $f(\tilde{\mathbf{w}}|\mathcal{M})$ by up-sweeping from leaves nodes to the current node in a single sub-band branch [30]. Moreover, likelihoods of all dyadic squares in the image can be obtained simultaneously in a single operation along the tree as well as Hidden Markov Tree (HMT) and Discrete Wavelet Transform (DWT) utilizes same dyadic squares in the quad-tree structure.

To obtain the likelihood of a sub-tree $\mathcal{T}_i$ of wavelet coefficients rooted at w_i, we have deployed wavelet HMT trees and learn parameters Θ for multiple levels [30]. The conditional likelihood $\beta_i(m) := f(\mathcal{T}_i|S_i = m, \Theta)$ can be retrieved by sweeping up to node i (see [30]); then, the likelihood of the coefficients in $\mathcal{T}_i$ can be computed as follows.

$$f(\mathcal{T}|\Theta) = \sum_{m=S,L} \beta_i(c)p(S_i = c|\Theta) \tag{4}$$

with $p(S_i = c|\Theta)$ state probabilities can be predefined or obtained during traning [31].

Due to similarity between wavelet coefficients sub-trees and dyadic squares, it is obvious that pixels of each square block $\mathbf{d_i}$ are well represented by three sub-bands or sub-trees $\{\mathcal{T}_i^{LH}, \mathcal{T}_i^{HL}, \mathcal{T}_i^{HH}\}$, whereof all likelihood are independently calculated by equation 4 in their corresponding trees . Independent estimation of three bands is an appropriate computation step because of the assumption that correlation between feature channels would not affect discriminant powers Gao [14]. Furthermore, DWT is known as de-correlation tools as well, and decorrelated signals are linearly independent from each other. Hence, the likelihood of a dyadic square is formulated as product of three independent likelihoods of wavelet sub-bands at each scale.

$$f(d_i|\mathcal{M}) = f(\mathcal{T}_i^{LH}|\Theta^{LH})f(\mathcal{T}_i^{HL}|\Theta^{HL})f(\mathcal{T}_i^{HH}|\Theta^{HH}) \qquad (5)$$

Noteworthy that, assumption of independent sub-bands are only necessary in the (U/T) HMT model while VHMT has grouped all coefficients into a single vector. Therefore, it only needs one quad-tree representation of multivariate coefficients, $\vec{\mathcal{T}}$, and the likelihood of dyadic squares under each tree node is formulated as follows.

$$f(d_i|\mathcal{M}) = f(\vec{\mathcal{T}}_i|\Theta) \qquad (6)$$

Using the equation 5, the likelihood can be computed for each dyadic square down to 2x2 block scale. Noteworthy, sub-band LL of the wavelet transform is not utilized in our computation since it is low-passed approximation of original images. Therefore, it is vulnerable to pixel brightness and lightning conditions of scenes. As natural images are considered as main testing category, shades and extreme brightness conditions can not be avoided; then, the final LL sub-band is discarded. The above simple formulation of likelihood is usually employed in block-by-block or "raw" classification since it does not exploit any possible relationship at different scales. Moreover, classification decisions between classes (centre and surround) are lack of inheritance across dyadic scales because a process of likelihood estimation at a scale is isolated from processes at other levels. Therefore, a better classifier can be achieved by integrating prior knowledge of other scales or at least the direct coarser scale.

4.5. Multi-scale Maximum a Posterior

In the previous section, only "raw" binary classification between two states have been realized under the wavelet Hidden Markov Tree model [30]. Given a prior knowledge from other scales, implementation of better binary

classification for DIS and MDIS, in the equation 2, needs a posterior probability $p(c_i^j|d^j)$ whereof c_i^j and $d^j = d_i^j$ are class labels and features of an image at a dyadic scale j and location i.

In order to estimate the MAP $p(c_i^j|d^j)$, we need to employ Bayesian approach for capturing dependencies between dyadic squares at different scales. Though many approximation techniques [36],[37],[38],[39] are derived for a practical MAP, the Hidden Markov Tree (HMT) by Choi [40] is proven to be a feasible solution. Choi [40] introduces hidden label tree modelling instead of joint probability estimation in high-dimensional data of dyadic squares. Due to strong correlation between the square under inspection and its parents and their neighbours, the decisions of class labels for these adjacent squares would affect the decision at the considered square. For example, if the parent square belong to a certain class and so do their neighbours, the child square most likely belongs to the same class as well.

Modelling the parent-child relation is realized by a general probabilistic graph [30]; however, the complexity exponentially increases with number of neighbouring nodes. Choi [41] proposes an alternative simpler solution, based on context-based Bayesian approach. For the sake of simplicity, causal contexts are only defined by states of the direct parent node and its 8 intermediate neighbours. Let's denote the context for D_i as $\mathbf{v_i} \equiv [v_{i,0}, v_{i,1}, \ldots, v_{i,8}]$ where $v_{i,0}$ refers to context from a direct parent node and their neighbours. The triple $\mathbf{v_i} \to C_i \to \mathbf{D_i}$ forms a Markov-1 chain, relating prior context $\mathbf{v_i}$ and node features $\mathbf{D_i}$ to classification decisions C_i. Moreover, class labels of prior contexts $\mathbf{v_i}$ are chosen as discrete values as it simplifies the modelling considerably. Given that prior context, independence can be assumed for label classification at each node; therefore, it is allowed to write.

$$p(\mathbf{c^j}|\mathbf{v^j}) = \prod_i p(c_i^j)|\mathbf{v_i^j}$$

The property of Markov-1 chain assumes that $\mathbf{D_i}$ is independent from $\mathbf{V_i}$ given C_i; therefore, the posterior probability of classifying $\mathbf{c^j}$ given $\mathbf{d^j}, \mathbf{v^j}$ is written as follows.

$$p(\mathbf{c^j}|\mathbf{d^j}, \mathbf{v^j}) = \frac{f(\mathbf{d^j}|\mathbf{c^j})p(\mathbf{c^j}|\mathbf{v^j})}{f(\mathbf{d^j}|\mathbf{v^j})}$$

As independence is assumed for label decisions in classifying processes, it yields.

$$p(\mathbf{c^j}|\mathbf{d^j}, \mathbf{v^j}) = \frac{1}{f(\mathbf{d^j}|\mathbf{v^j})} \prod_i f(\mathbf{d^j}|c^j)p(c^j|\mathbf{v^j})$$

21

and the marginalized context-based posterior

$$f(c_i^j|\mathbf{d^j}, \mathbf{v^j}) \propto f(\mathbf{d_i^j}|c_i^j)p(c_i^j|\mathbf{v_i^j})$$

It greatly simplifies MAP posterior estimation since it no longer needs to deal with joint prior conditions of features and contexts. It only needs to obtain two separated likelihood of the dyadic square given the class value C_i, $f(\mathbf{d_i}^j|c_i^j)$ and prior context provided through $\mathbf{v_i}$, $p(c_i^j|\mathbf{v_i^j})$.

While retrieving the likelihood $f(\mathbf{d_i}^j|c_i^j)$ is straightforward by up-sweeping operations with given HMT model parameters at each scale, the complexity of prior context estimation greatly depends on its structures. Though more general structures may give better prior information for classification, it also greatly complicates the modelling and summarizing information conveyed by $\mathbf{v_i^j}$ as well. In other words, we run on the verge of context dilution, especially in case of insufficient training data [36],[37],[30].

To simplify but still guarantee generalization of prior information, we will employ a simple context structure inspired by the hybrid tree model [39] in context-labelling trees. Instead of including all neighbouring sub-squares, the simplified context only involves labels from the parent square $C_{\rho(i)}$ and major vote of the class labels from neighbouring squares $C_{\mathcal{N}_i}$. As there are only two class labels $N_c := 0, 1$, the prior context $\mathbf{v_i} := \{C_{\rho(i)}, C_{\mathcal{N}_i}\}$ can only been drawn from $N_c^2 = 4$ different values $0, 0, 0, 1, 1, 0, 1, 1$. Despite such ad-hoc simple contextual model, it provides sufficient statistic for demonstrating the effectiveness of multi-scale decision fusion [39]. Another advantage of the context structure simplification is not requiring enormous number of training data for probability estimation.

Any decision about labels at a scale j depends on prior information of labels on a scale $j - 1$; therefore, we can maximize MAP 7 in multi-scale coarse-to-fine manner by fusing the likelihoods $f(\mathbf{d_i}|c_i)$ given the label tree prior $p(c_i^j|\mathbf{v_i})$. The fusion step help pass down MAP estimation through scales to enhance coherency between classifying results of consecutive scales. Moreover, a posterior probability of a class label c_i given features and the prior context is computed and maximized coherently across multiple scale.

$$\hat{c}_i^{MAP} = argmax_{c_i^j \in 0,1} f(c_i^j|\mathbf{d^j}, \mathbf{v^j}) \tag{7}$$

4.6. Multiscale Discriminant Saliency

Core ideas of DIS and MDIS are measuring discriminant power between two classes centre and surrounds. Though it can be estimated by sample

22

means of mutual information, the underlying mechanism is distinguishing the centre and surround classes given Generalized Gaussian Distribution (GGD) of wavelet coefficients. These distributions are usually zero-mean and well-characterized with only variance parameters. Dashan Gao [14] has estimated scale parameter (variance) of GGD (see section 2.4 [14] for more details) by the maximum a probability process.

$$\hat{\alpha}^{MAP} = \left[\frac{1}{\mathcal{K}}\left(\sum_{j=1}^{n}|x(j)|^{\beta} + \nu\right)\right]^{\frac{1}{\beta}}$$

The above MAP is later used for deciding whether a sample point or an image data point belongs to either the centre or surround class (see [14] for a detailed proof and explanation). Then, the more distinguishing MAP estimation of the centre class's variance parameter α_1 is from that of the surround class's variance parameter α_0, there is more discriminant power for classifying interest from null hypothesis.

Beside GGD, Gaussian Mixture Model (GSM) is a popular choice for modelling wavelet distributions with multiple classes' variances as well [33],[34],[42],[35]. In binary classification problem with only two classes, GSM includes two Gaussian Distribution mixtures (GD) of distinguishing variances, which are named as "large" / "small" states according to their comparison in terms of variance values. Now the only difference between GSM models and Gao's proposal [13] are whether GD or GGD should be used. Though GGD is more sophisticated with customizable distribution shape parameter β, several factors support validity of simple GD modelling given the class conditions as hidden variables. Empirical results from estimation have shown that the mixture model is simple yet effective [33],[39]. Modelling wavelet coefficients with hidden classes of "large"/"small" variance states are basic data models in Wavelet-based Hidden Markov Model (HMT) [30]. With wavelet HMT, image data are processed in coarse-to-fine multi-scale manner; therefore, MAP of a state C_i^j given input features from a sub-square D_i^j can be inherently estimated across scales $j = 0, 1, \ldots, J$. More details about this multi-scale MAP estimation by wavelet HMT can be found in the previous sections 4.5. Then, a combination of MAP estimation, in the equation 7 and mutual information computation, in the equation 2, yields the MDIS mathematical formulation.

$$I_i^j(C^j; \mathbf{D^j}) = H(C^j) + \sum_{c=0}^{1} P_{C^j|\mathbf{D^j}}(c_i^j|\mathbf{d^j}) log P_{C^j|\mathbf{D^j}}(c_i^j|\mathbf{d^j}) \qquad (8)$$

23

where $H(C^j) = -p(C^j)log(p(C^j))$ is entropy estimation of classes across scales j, and the posterior probability can be estimated by modelling wavelet coefficients in the HMT framework. This matter has been discussed in previous sections; therefore, it is not repeated here. As the equation 8 yields discriminant power across multiple scales; a strategy is needed for combining them across scales. In this paper, a simple maximum rule is applied for selecting discriminant values from multiple scales into a singular MDIS saliency map at each sub-square d_i.

$$I_i(C|\mathbf{D}) = max\left(I_i^j(C^j; \mathbf{D^j})\right) \tag{9}$$

5. Experiments & Discussion

In the light of decision theory, saliency maps are considered as binary filters applied, at each image location. According to a certain saliency threshold, each spot can be labelled as interesting or uninteresting. If a binary classification map is considered as saliency map which leads to visual performance of human beings, it would be a significant undervaluation of human visual attention system. Psychology experiments shows much better capacity of biologically plausible visual attention system than that of binary classification maps. Therefore, our proposed method just use binary classification between centre-surround environment as an intermediate step to develop information-based saliency map. At each location, mutual information between distributions of classes and features shows strength of discriminant power between interesting vs non-interesting classes given input dyadic sub-squares. From discrete binary values, the saliency representation has continuous ranges of discriminant power. Inevitably, the generated saliency maps become more correlated to the results of human visual attention maps.

Besides suitable saliency representation, reliable ground truth data are necessary for evaluation. As our research purpose is deepening knowledge about multi-scale discriminant saliency approaches and human visual attention relation, the ground truth data must be gotten from psychological experiments in which human subjects look different natural scenes and their responses are collectively acquired. Moreover, the research scope only focus at bottom-up visual saliency, the early stage of attention without interference of prior knowledge and experiences. Human participants should be naive about aims of experiments and should not know contents of displaying scenes in advance. After these prerequisites are satisfied, human responses

on each scene can be accurately collected through eye-tracking equipments. It records collection of eye-fixations for each scene, and these raw data are basic form of ground truths for evaluating efficiency of saliency methods.

Assumed that ground-truth data are available, quantitative methods can be applied for evaluation of MDIS Saliency results. In an effort of standardizing evaluating process, we only utilize one of the most common and accessible database and evaluation tools in visual attention fields. With regard to available database and ground-truths, Niel Bruce database [9] is certainly the most popular dataset used in information-based saliency studies. While proposing his InfoMax (AIM) approach, the first information-based visual saliency, he simultaneously releases his testing database as well. The reasonably small collection with 120 different colour images which are tested by 20 different subjects. Each object observes displayed images in random orders on a 12 inch CRT monitor positioned 0.75 from their location for 4 seconds with a mask between each pair of images. Importantly, no particular instructions are given except observing the image.

Above brief description clarified validity of this database for our experiments. Besides that, the AIM method is involved as a reference method against which we compare our proposed saliency solution MDIS in terms of performance, computational load, etc. Though DIS [13] is the closest approach to our proposed MDIS, implementation from the author is not available for comparison. Meanwhile, AIM also derives saliency value from information theory with slightly different computation, self-information instead of mutual-information in MDIS or DIS. Therefore, it would be considered the second best as referenced method for our later evaluation of MDIS.

As valid database is set, proper numerical tools are necessary for analysing simulation data. Regarding fairness and accuracy of the evaluation, we employ a set of three measurements LCC, NSS, and AUC recommended by Ali Borji et al. [43] since evaluation codes can be retrieved freely from their website [2]. Assessment with three evaluation scores ensures the reliability of qualitative observation and any conclusion is free from metric choices. First, linear correlation coefficient (LCC) measures linear relationship between two variables $CC(G,S) = cov(G,S)$, where G and S are the standard deviation of a ground-truth and saliency maps. Values of LCC variates from -1 to +1 while the correlation changes from total inverse to perfect linear relation

---

[2]https://sites.google.com/site/saliencyevaluation/

between G and S maps.

While LCC measures similarity of saliency and fixation maps as a whole, normalized scan path saliency (NSS) treats eye-fixations as random variables which are classified by proposed saliency maps. The measurement is an average of saliency values at human eye positions according to each approach. NSS metrics are ranged from 0 to 1; $NSS = 1$ indicates that saliency values at eye-fixation locations are one standard deviation above average; meanwhile, $NSS = 0$ means no better performance of saliency maps than randomly generated maps.

Different from previous two measurements dealing with saliency maps directly, the last quantitative tool AUC utilized saliency maps as binary classification filters with various thresholds. By regularly changing threshold values over the range of saliency values, we can have a number of binary classifying filters. Then, deploying these filters on eye-fixation maps produce several true positive rates (TPR) and false positive rate (FPR) as vertical and horizontal values of Receiver Operating Characteristics (ROC) curve. Area under curve (AUC) is a simple quantitative measurement to compare ROC of different saliency approaches. Perfect AUC prediction means a score of 1 while 0.5 indicates by chance-performance level. In addition to ROC, ISROC plots [7] are useful for clarifying performances of evaluated approaches against those of multiple human experiment subjects.

As mentioned previously, AIM is chosen as the referenced information-based saliency method. It is chosen due to the well-established reputation as well as freely accessible code and experiment database [3]. Due to multiscale natures of the proposed MDIS, saliency maps for each dyadic-scale levels can be extracted as well as the final MDIS saliency, integrated across scales by the equation 8. The availability of saliency maps at multiple scales as well as combined one allows evaluation of discriminant power concept for saliency in scale-by-scale manner or on the whole. We denote integrated MDIS as HMT0; while separated saliency maps are named as HMT1 to HMT5 in accordance with coarse-to-fine order. By examining MDIS in different aspects, we would observe its effectiveness in predictions of eye-fixation points and how selection of classifying window sizes might affect its performance. In addition, comparisons against AIM would contribute a general view how MDIS performs against a well-known information-based saliency method.

---

[3]http://www-sop.inria.fr/members/Neil.Bruce/

In our proposed methods, Hidden Markov Tree (HMT) plays a role of modelling statistical properties of images. It extracts model parameters at each scale by considering distribution of feature given hidden variables (centre-surround labels in our method). Therefore, training is a necessary step before model parameters can be approximated in THMT. Normally different wavelet sub-bands are trained independently assumed no-correlation exists between them. However, HMT parameters can be acquired by grouping wavelet coefficients and training these multivariate variables in VHMT models as well. Furthermore, training steps are not necessarily needed if input images are all from a specific category; for example, natural images in this paper. Romberg et al. [31] have studied this Universal Hidden Markov Tree (UHMT) for the natural image class. In other words, model parameters can be fixed without any training efforts; the approach would greatly reduce computational load for MDIS. Each configuration of the HMT model has its own advantages and disadvantages; therefore, a simulation is necessary to compare performances of (U/T/V) HMTs in terms of LCC, NSS, AUC and TIME (the computational time). Noteworthy that, we will employ (U/T/V) HMTs instead of HMT only when representing experimental data in order to signify which tree-building method is employed.

In quantitative method, general ideas can be drawn about how the proposed algorithms perform in average. However, such evaluation method lacks of specific details about successful and failure cases. The information has been averaged out in quantitative method. In an effort of looking for pros and cons of the algorithm, we perform qualitative evaluations for saliency maps generated by MDIS in multiple scales. Furthermore, AIM saliency maps are generated and compared with MDIS maps to discover advantages and disadvantages of each method.

5.1. Quantitative Evaluation

After general review of how simulations are built and evaluated in the previous section, following are data representation and analysis of the conducted experiments. In this paper, five dyadic scales are deployed for any HMT training and evaluation; therefore, we have simulation modes from (U/T/V)HMT(1-5) of MDIS depending on whether training stages is deployed (T/V)HMT or universal parameters are used (UHMT). Saliency maps could be combined according to the maximization of mutual information rule, the equation 8; therefore, we have three (U,T,V)HMT0 modes for saliency maps which are created by across-scale integration. AIM is involved in the

simulations as the reference method; meanwhile, LCC, NSS, AUC and TIME are chosen as numerical evaluation tools. Below are two tables of simulation results. Table 1 shows experimental data of all universal HMT modes while table 2 summarizes data of all trained HMT modes.

Table 1: UHMT - MDIS - DATA

Observations	LCC	NSS	AUC	TIM
UHMT0	0.01434	0.21811	**0.89392**	0.39617
UHMT1	-0.00269	0.19772	0.53862	0.39617
UHMT2	0.01294	0.27819	0.60520	0.39617
UHMT3	0.01349	0.32868	0.69065	0.39617
UHMT4	**0.01604**	**0.42419**	0.83615	0.39617
UHMT5	0.00548	0.13273	**0.89234**	0.39706
AIM	0.01576	0.12378	0.72275	50.41714

Table 2: THMT - MDIS - DATA

Observations	LCC	NSS	AUC	TIM
THMT0	**0.02382**	**0.48019**	**0.88357**	2.32734
THMT1	**0.02582**	0.38096	0.60922	2.32734
THMT2	0.01156	0.31855	0.64633	2.32726
THMT3	0.01604	0.32491	0.71972	2.32726
THMT4	0.01143	0.29662	0.81192	2.32726
THMT5	0.00512	0.36932	**0.89532**	2.32726
AIM	0.01576	0.12378	0.72353	50.41714

Table 3: VHMT - MDIS - DATA

Observations	LCC	NSS	AUC	TIM
VHMT0	0.01697	**0.44170**	**0.86606**	2.84212
VHMT1	0.01693	0.38387	0.61187	2.84212
VHMT2	**0.02044**	0.38777	0.67060	2.84212
VHMT3	0.01430	0.38882	0.73682	2.84212
VHMT4	0.00946	0.36761	0.82329	2.84212
VHMT5	-0.00125	0.39580	**0.88160**	2.84212
AIM	0.01576	0.12378	0.72400	50.41714

Among numerical evaluation tools of visual saliency, Receiver Operating Curve (ROC) and its Area Under Curve (AUC) are the most popular. It measures efficiency of saliency maps in classifying fixation and non-fixation points of human eye movements in visual psychological experiments. In ROC curve, the vertical axis indicates True Positive Rate of the classification which is equal to hit rate, recall measurement. It is the ratio between correctly classified fixation points and its total number. Meanwhile, False Positive Rate (FPR) is equivalent with fall-out ratio, the number of incorrectly classified
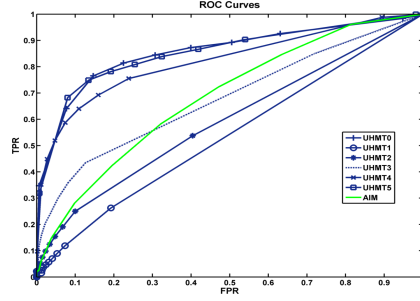
(a) UHMT - MDIS - ROC
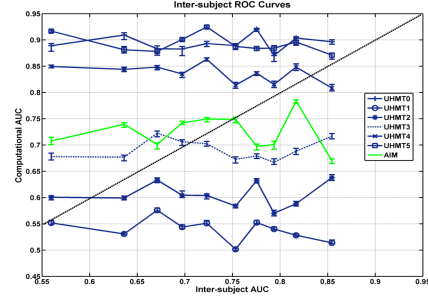
(b) UHMT - MDIS - ISROC
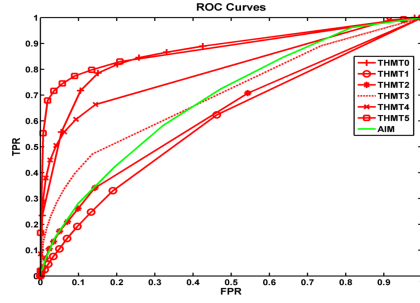
(c) THMT - MDIS - ROC
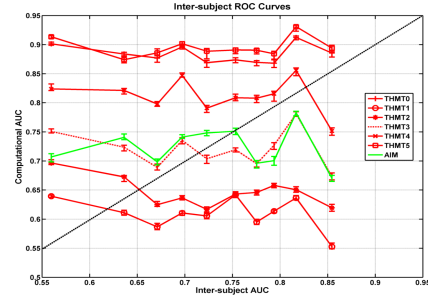
(d) THMT - MDIS - ISROC
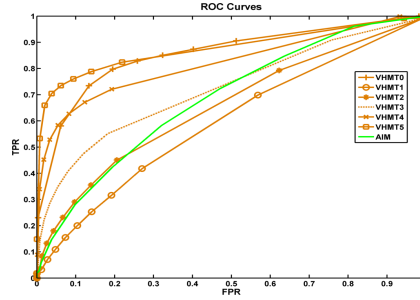
(e) VHMT - MDIS - ROC
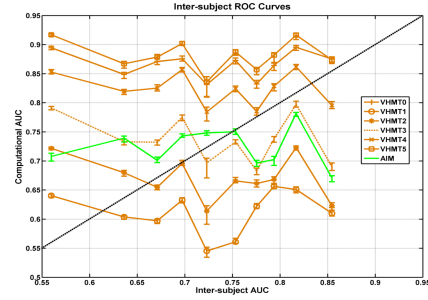
(f) VHMT - MDIS - ISROC

Figure 3: (U/T/V)HMT - MDIS - (IS)ROC

fixation points over the total number of non-fixation points. In brief, ROC curves represent a successful rate (TPR) of guessing eye-fixation points at a particular falling rate (FPR) over a normalized continuous range of threshold from 0 to 1. When a threshold is raised up to the maximum saliency value, there are no interested points. Then, it is the case when both hit and fall rates are 0, $(FPR, TPR) = (0, 0)$. When the bar is lowered gradually, the re-

call and fall-out rates increase bit by bit as well. If the recall increases faster than the fall-out rate, a binary classification performs well since it tends to make a correct decision than wrong decisions about which locations are interesting. Otherwise, it is unreliable in the choice of salient points. Until the minimum threshold is reached, both recall and falling out rates becomes maximum, $(FPR, TPR) = (1, 1)$, the case of no interesting points.

Figures 3(a),3(c),3(e) display ROC curves of UHMT, THMT, VHMT for several scales with AIM as reference curves. In all figures, solid green lines are representing ROCs of AIM, the reference method; while, blue, red, and orange colours represent ROCs of UHMT, THMT, and VHMT consequently. In general, AIM has an average performance in comparisons of other HMT with different scales and window sizes. It performs better than all HMT when large 16x16 or 32x32 windows are utilized. When 8x8, 4x4, and 2x2 blocks are employed, HMTs surpass AIM in detection of fixation points. In other words, HMT with these scales produces more meaningful maps than AIM does. Noteworthy that AIM utilizes 30x30 windows, almost equivalent with (U/T/V) HMT5's window sizes; however, AIM generates overlapping squares instead of distinct blocks. Utilization of such large windows and high-dimensional data requires training steps and samples preparation which require enormous computational effort. Application of machine learning in continuously sliding windows explains superiority of AIM over (U/T/V) HMT with similar block-sizes. However, performance of HMTs in terms of ROC shapes rapidly over-perform AIM if we shrink block sizes from HMT1 to HMT5, in other words, increase the resolution of output saliency maps.

In comparisons of (U,T,V) HMTs, we can see advantages of (T,V)HMT over UHMT on different scales. It is clearly shown in three corresponding ROC figures 3(a),3(c), 3(e) since curves of (T,V)HMT moves closer toward the left-top corner than UHMT does. It means (T,V)HMT more successfully detect meaningful points than UHTM does in region of low fall-out rate (FPR). Therefore, (T,V)HMT provides better binary classifier in term of robustness. It is reasonable since THMT, VHMT requires training steps on image data meanwhile UHMT model just uses predefined and general parameters. Furthermore, VHMT has slightly better performance than THMT in most window sizes. The slight increment in terms of ROC curves is due to the fact VHMT is a better model than THMT in processing texture features [27].

Generalized ROC evaluates performances of saliency maps with different

thresholds, according to ground-truth eye-fixation data of several human test subjects. However, it does not distinguish eye-fixation data from each subject but treats the data collectively. Therefore, it certainly loses important aspects of eye-tracking data from multiple subjects such as diversity of individual responses over various types of scenes shown during experiments. For example, a simple scene with only one simple subject on the blank white background would cause similar eye-movement from all subjects. However, these responses are rather diverge when scenes with two or more foreground subjects on complex experiments are shown to participants. To visualize the diversity of subjects' responses and corresponding performances of saliency maps, we utilize evaluation method similar to Inter-Subject ROC approach of Harel .et .al [7], whereof saliency methods are evaluated against diversity in subjects' responses. The horizontal axis, inter-subject AUC, averagely measures how similarly an eye-fixation map of a subject resembles the map generated by eye-tracking data from the other subjects. If the value is small, it means not much consistency across subjects' performance due to complex scenes. Otherwise, it shows that simple scenes results in agreement of responses from most subjects.

Applying the ISROC evaluation method for the proposed HMTs and the reference AIM saliency methods generates figures 3(b),3(d), and 3(f) for (U) universal, (T) trained and vector-based configuration. Generally, all saliency methods have quite consistent performances across different complexity of scenes. In other words, they could produce meaningful saliency maps in complex cases where inter-subject scores are low, human subjects are struggling to find common salient points. However, these computational saliency maps are overcomplicate in simple scenes with high inter-subject ROC scores . In these cases, all subjects focus on a few locations on testing scenes, while the proposed saliency methods still detect other regions of interest beside the main objects. Therefore, the proposed computational approaches are less efficient than human beings in such situations.

Similar to ROC curves, ISROC curves help to compare HMTs with different configurations on multiple levels as well. Observing figure 3 shows advantages of HMTs in terms of ISROC curves when smaller block sizes are chosen. Moreover, the reference method AIM's performance, the solid green line, is better than HMT0 (16x16 blocks) and HMT1 (32x32 blocks), equivalent with HMT3 (8x8 blocks), but worse than HMT0, HMT4 (2x2 blocks), and HMT5 (4x4 blocks). HMT0, HMT4 and HMT5 have better performance than human subjects in most of situation; their ISROC curves are above the

black boundary set by human subjects for most of the plots. Noteworthy, there is an interesting observation about HMT0 - the integrated map, and HMT5 - 2x2 blocks that HMT5 is almost equivalent with or sometimes better than the integrated method HMT0. Perhaps, more complex rules for integrating saliency maps need developing to efficiently integrate attention maps with different scales.

Shapes of ROC and ISROC curves give general ideas that HMTs overperform the reference AIM map; however, it does not specify how much better the proposed methods are. Hence, it is necessary to have numerical analysis of their performances. Computational loads are the first measure to be analysed and compared among saliency approaches. In the TIME rows of tables 1,reftab:thmt and 1, we present necessary processing time for each method or each mode. Generally, computational loads, proportional to processing time, of all modes in either UHMT or THMT row are almost similar since the parameters of full-depth Hidden Markov Tree need estimating before computation of saliency values. In comparison of (U/T/V) HMTs in terms of processing time, UHMT is faster than (T/V)HMT as UHMT uses predefined parameters instead of learning HMT parameters from each image. When comparing (U/T/V) HMT modes of MDIS with AIM, our proposed methods are much faster than AIM. The well-known AIM directly estimates self information from high-dimensional by ICA algorithm while MDIS statistically models two hidden states: "large" "small" states in sparse and structural features. Computational load or processing time of the mentioned AIM and proposed MDIS with different modes can be seen in the figure 4(d). Though HMT-MDIS significantly reduces computational load for computation of information-based saliency, more verifications are necessary for their performances in terms of accuracy. We begin with evaluating three modes (U/T/V)HMT separately with AIM in terms of three numerical tools LCC, NSS, AUC together, Figure 4(a), 4(b), 4(c). Then all modes of HMTs are summarized in three plots (the top row of Figure 5) in the following order NSS, LCC, AUC from left to right. Especially in the figure 5, simulation modes of the same scale level are placed next to each other for example (U/T/V) HMT0 sit next to each other, so do (U/T/V) HMT1 and etc. It is intentionally arranged in that way to compare performances of different simulation modes in the same scale level. Noteworthy that, in both tables 1 and 2 for each row is identified **maximum** and `minimum` values by using corresponding text styles. Identification of extreme values only involves derivatives (U,T,V) HMT of MDIS modes. In the figures 4(a), 4(b), 4(c), extreme

(a) UHMT - MDIS
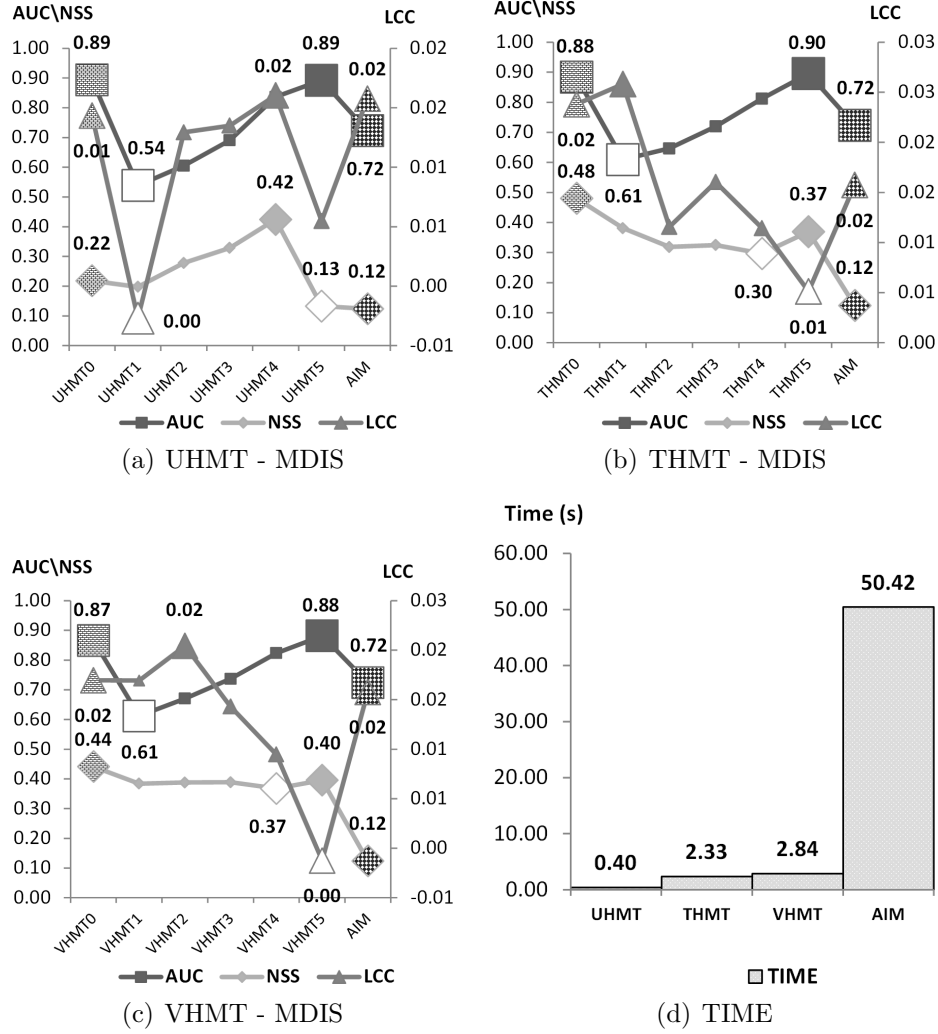
(b) THMT - MDIS

(c) VHMT - MDIS

(d) TIME

Figure 4: Performance of UHMT-MDIS, THMT-MDIS in AUC,NSS,LCC and TIME

(maximum or minimum) values are also specially marked. For example, maximum values have big solid markers while big but empty ones represent minimum points. Especially, AIM has big markers with distinguishing big cross-board texture while integrated saliency modes of (U/T/V)HMT0 have small cross-board textures. These special markers help highlight interesting facts in comparison among HMTs-MDIS or MDIS against AIM. The same marking policy is applied for data representation in the figure 5. Meanwhile,

each line in this figure has an arrow head for showing trends of experimental data (increasing/decreasing) when simulation modes are changed across U,T, or V configurations of HMT for each scale level.
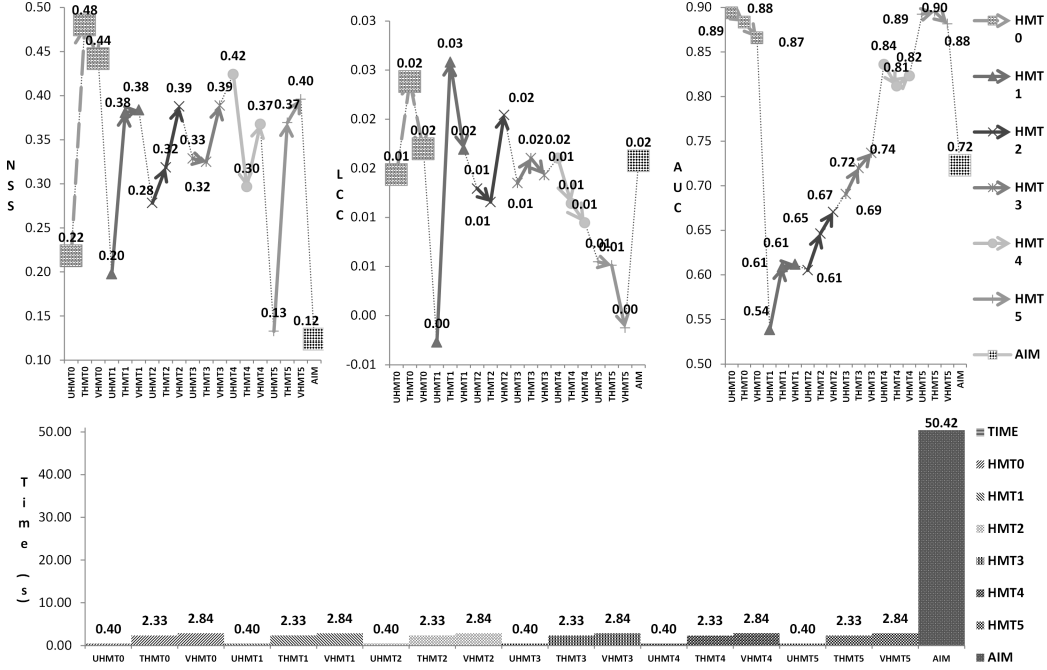
Figure 5: Summary of all MDIS against AIM

Firstly, the MDIS approach with universal parameters for each level of hidden Markov tree is analysed in terms of accuracy since it requires very little effort in saliency computation. Obviously, that fact raises a question about its accuracy of centre and surround classifier as well as synthesized saliency maps. According to simulation data in the table 1 with highlighted extrema, UHMT performs pretty well against AIM in all three measurements LCC, NSS and AUC. For example, MDIS with UHMT4 mode (4x4 square blocks) surpasses AIM in all measurements. It confirms validity and efficiency of our proposed methods in the information-based saliency map research field. When performances of different UHMT-MDIS modes are considered, UHMT4 with 4x4 squares have the most consistent evaluation among all dyadic scales with maximum LCC and NSS and the second best AUC value. UHMT0-MDIS, integration of saliency values across scales, does not have better performance than other UHMTs except for AUC level. It

34

shows inconsistent side of deploying HMT with predefined universal parameters while no training effort is done for adapting the model into multi-scale statistical structures.

Secondly, training stages are included in the simulation of MDIS with (T/V)HMT mode (Trained Hidden Markov Tree). With additional adaptivity, (T/V)HMT might improve the saliency evaluation and produce more consistent results than UHMT might. This subjection is solidified by simulation data in the tables 2,3 and they are also plotted in the figures 4(b),4(c). As observed in the table, all **maximum** values locate at the THMT0 column, THMT0-MDIS over-performs AIM in all evaluating schemes. Again, the rationale of MDIS is confirmed and practically proved. Furthermore, effectiveness of training stages are clearly shown when comparing THMT0 against UHMT0. Though AUC of THMT0 is smaller than that of UHMT0, THMT0 evaluation are better their counterparts in both NSS and LCC schemes. This confirms usefulness of training Hidden Markov Tree models for each sample image. In addition, the figure 4(b),4(c) shows supremacy of (T,V) HMT0 modes, the across-scale integration mode of MDIS over other singular saliency maps at different dyadic scales in any measurement. Noteworthy, that LCC of THMT0 mode is a bit smaller than LCC of THMT1 mode; however, this small difference can be safely ignored. Comparison of UHMT-MDIS and THMT-MDIS mode-by-mode between data in Table 4(a) and Table 4(b) are shown in the figure 5. Accordingly, there are slight improvements of (T/V)HMT1,(T/V)HMT2 over UHMT1, UHMT2, equivalence of THMT3, UHMT3, and a reverse trend that UHMT4-5 are comparable or slightly better than THMT4, THMT5. It seems that training processes are more important when big classification windows are used. Meanwhile universal approaches of HMT work pretty well if dyadic squares get smaller. Two possible reasons for this observation are statistical natures of dyadic squares and characteristics of training processes. A bigger square has richer joint-distribution of features; therefore, UHMT with fixed parameters can not marginally approximate that distribution well. However, (T,V)HMT parameters models can be learn from analysing images; it results in significant improvement of quality of saliency maps. While smaller sub-squares are less statistically distinguishing, they are successfully modelled by universal parameters of HMT. Then training processes might become redundant since UHMT would perform as well as THMT would do.
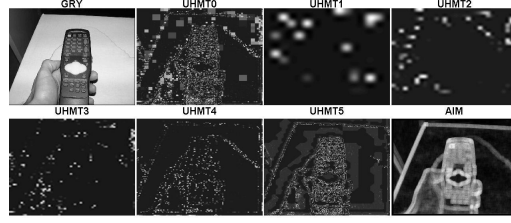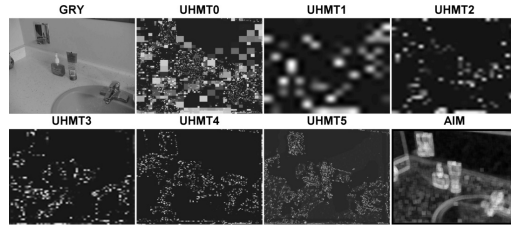
(a) UHMT
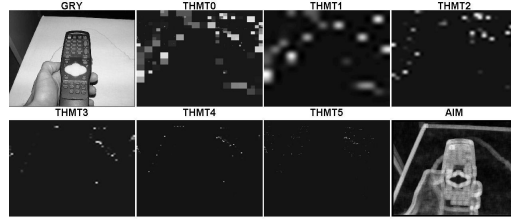
(b) THMT
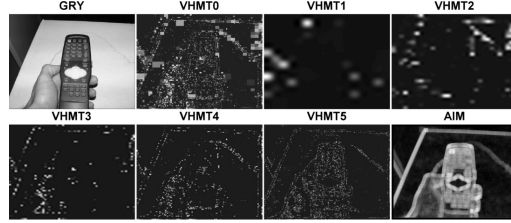
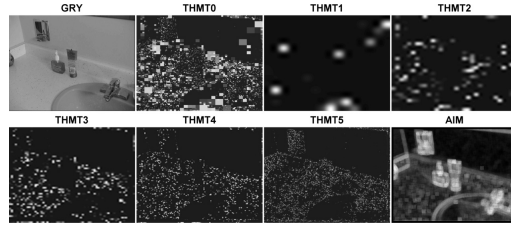
(c) VHMT

Figure 6: Saliency Maps 1A

5.2. Qualitative Evaluation

In this section, saliency maps are analyzed qualitatively or visually. From this analysis, we want to identify (i) on which image contexts (U/T/V) HMT-MDIS work well, (ii) how scale parameters affect formation of saliency maps, and (iii) how MDIS in general is compared with AIM.

In figures 6 7, the first example with central objects shows an example of good (U/V)HMT performance but bad THMT performance. All scale levels of THMT suppress features of the most obvious objects in the image center. Meanwhile, (U/V)HMT4 and (U/V)HMT5 capture significant features points of that objects; therefore, (U/V)HMT0 have much better saliency map than THMT0. In this case, the best saliency map of MDIS approach, (U/V)HMT0, is reasonably competitive against AIM one. Despite different

36

(a) UHMT

(b) THMT
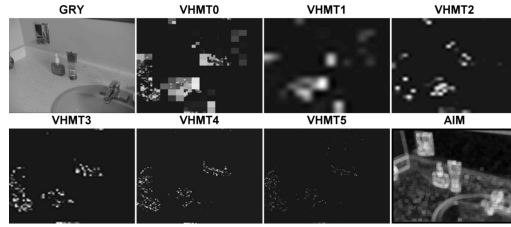
(c) VHMT

Figure 7: Saliency Maps 1B

configurations, UHMT-MDIS and VHMT-MDIS have quite similar results for a simple scenes with few central objects. Observing the figures 6(a),7(a) and 6(c),7(c), we can see differences between generated saliency maps of UHMT and VHMT across six different modes [0-5] especially mode 5. This mode utilizes the smallest 2x2 blocks; therefore, it can create saliency maps with great details about discrepant regions between (U/V)HMT maps . Generally, UHMT tends to include more irrelevant areas than VHMT does; while VHMT only focuses on regions richer of edges and textures. It is coherent with quantitative comparisons between UHMT and VHMT in the tables 1, 3. However, there exists exceptions like the case in figure 7 where VHMT misses edges and texture of foreground objects but wrongly focus on small black circles on the background. Though universal mode does not have on-
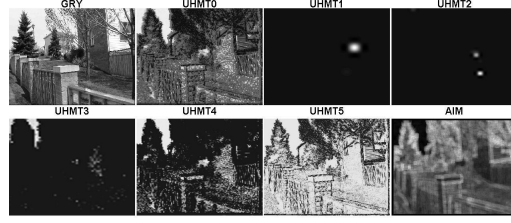
(a) UHMT
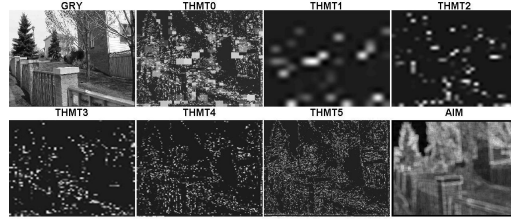
(b) THMT
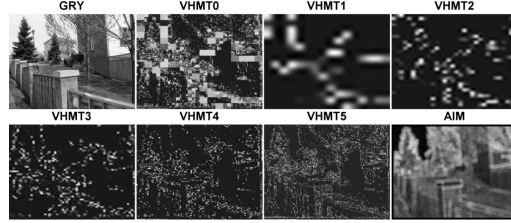
(c) VHMT

Figure 8: Saliency Maps 2A

line training, the predefined models are learnt by off-line training of many images. Therefore, it robustly performs in many cases which others fail to operate well. Meanwhile, saliency maps generated by VHMT-MDIS are much attracted to small black dots on background due to rotational invariance nature of the vector-HMT model [27].

Different from previous examples, the figures 8,9 show an opposite case in general outdoor scenes which (T/V)HMT produce more reasonable saliency maps for. While UHMT0 map covers the whole region of sky despite no interesting features, (T/V)HMT0 correctly focus on interesting objects on the scene. Similarly, (V/T)HMT does extract more meaningful features than UHMT does (see the figures 8,9). In addition, the best saliency map THMT0 or THMT5 highlights more discriminant features than AIM saliency map.
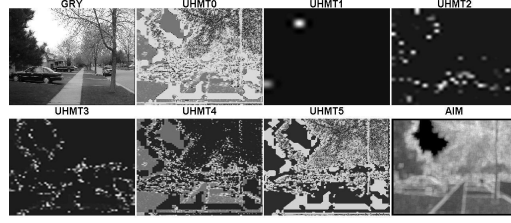
(a) UHMT
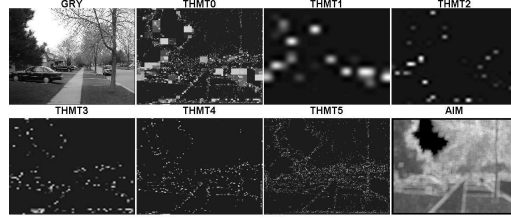
(b) THMT
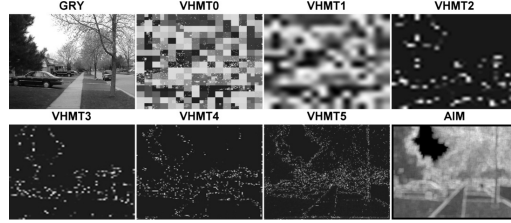
(c) VHMT

Figure 9: Saliency Maps 2B

Noteworthy, there are significant differences between VHMT1 and THMT1 saliency maps. While THMT1 over-emphasizes edge points and lines between appeared textures such as trees and the sky, the VHMT1 is ridged with highlighted regions and does not hint any standing-out areas. Again this discrepancy in performance may be due to nature of orientation selection in each mode (T/V) as THMT favours only features of horizontal, vertical, diagonal directions; while VHMT mode, a rotation invariant scheme, does not have any oriental differentiation. The third pair of examples are chosen such that complex scenes are presented to the saliency methods. In the figure 10, there are several fruits on the shelf; it is considerably complicated due to richness of edges, textures, as well as colour. In general, both (U/T/V)HMT-MDIS and AIM only partially succeed in detecting saliency
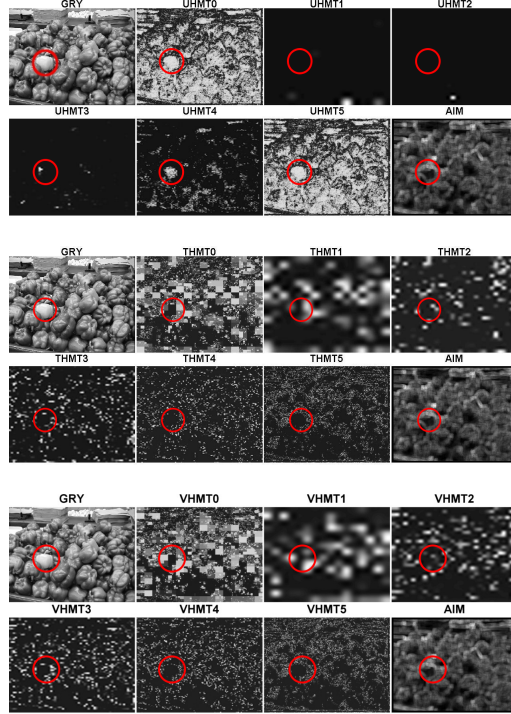
39

Figure 10: Saliency Maps 3A

regions from these images since none of them successfully highlight the fruit with different colour on the shelf (the fruit inside a red circle, Figure 10). Though most MDISs for variety of scale levels, do not explicitly detect that fruit, UHMT3 and UHMT4 salency maps are able to highlight the location of that fruit (see UHMT3 an UHMT4 saliency maps, Figure 10). The sample matches with the fact that UHMT4 data in the table 1 has extremely good performance in all evaluation schemes. Surprisingly, there are some cases when appropriate choices of scales and parameters of predefined HMT models can over-perform all trained HMT models. The interesting example of the figure 10 opens another research direction about how HMT model can be learned and optimized; however, it is the question of another research paper.

6. Conclusion

In conclusion, the multiple discriminant saliency (MDIS), a multi-scale extension of DIS [21] under dyadic scale framework, has strong theoretical

foundation as it is quantified by information theory and adapted to multiple dyadic-scale structures. The performance of MDIS against AIM is simulated on the standard database with well-established numerical tools; furthermore, simulation data proves competitiveness of MDIS over AIM in both accuracy and speed. However, MDIS fails to capture salient regions in a few complex scenes; therefore, the next research step are improving MDIS accuracy in such cases. In addition, implementation of MDIS algorithm in embedded systems is also considered as a possible research direction.

References

[1] U. Neisser, Cognitive psychology., APA.
URL http://psycnet.apa.org/psycinfo/1967-35031-000

[2] D. Marr, D. Marr, Early processing of visual information, Philosophical Transactions of the Royal Society of London. B, Biological Sciences 275 (942) (1976) 483–519.
URL http://rstb.royalsocietypublishing.org/content/275/942/483.short

[3] A. M. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive psychology 12 (1) (1980) 97–136.
URL http://www.sciencedirect.com/science/article/pii/0010028580900055

[4] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, Pattern Analysis and Machine Intelligence, IEEE Transactions on 20 (11) (1998) 1254–1259.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=730558

[5] D. Gao, N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes, Advances in neural information processing systems 17 (481-488) (2004) 1.
URL http://www.svcl.ucsd.edu/publications/conference/2004/nips04/nips04.pdf

[6] Y. Sun, R. Fisher, Object-based visual attention for computer vision, Artificial Intelligence 146 (1) (2003) 77–123.

URL   http://www.sciencedirect.com/science/article/pii/
S0004370202003995

[7] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, NIPS.
    URL http://papers.klab.caltech.edu/300/1/543.pdf

[8] X. Hou, L. Zhang, Saliency detection: A spectral residual approach,
    in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE
    Conference on, 2007, p. 1–8.
    URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=
    4270292

[9] N. Bruce, J. Tsotsos, Saliency based on information maximization, Ad-
    vances in neural information processing systems 18 (2006) 155.
    URL http://www.cse.yorku.ca/~tsotsos/Homepage%20of%20John%
    20K_files/NIPS2005_0081.pdf

[10] A. C. Le Ngo, G. Qiu, G. Underwood, L. M. Ang, K. P. Seng,
    Visual saliency based on fast nonparametric multidimensional entropy
    estimation, in: Acoustics, Speech and Signal Processing (ICASSP),
    2012 IEEE International Conference on, 2012, p. 1305–1308.
    URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=
    6288129

[11] A. C. Le Ngo, L. M. Ang, K. P. Seng, G. Qiu, Improvement and
    evaluation of visual saliency based on information theory, in: Computer
    Symposium (ICS), 2010 International, 2010, p. 500–505.
    URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=
    5685461

[12] G. Qiu, X. Gu, Z. Chen, Q. Chen, C. Wang, An information theoretic
    model of spatiotemporal visual saliency, in: 2007 IEEE International
    Conference on Multimedia and Expo, 2007, pp. 1806 –1809. doi:10.
    1109/ICME.2007.4285023.

[13] D. Gao, V. Mahadevan, N. Vasconcelos, The discriminant center-
    surround hypothesis for bottom-up saliency, Advances in neural
    information processing systems 20 (2007) 1–8.
    URL   http://132.239.134.221/publications/conference/2007/
    nips2007/nips2007_budiscsal.pdf

[14] D. Gao, N. Vasconcelos, Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics, Neural Computation 21 (1) (2009) 239–271.
URL http://www.mitpressjournals.org/doi/abs/10.1162/neco.2009.11-06-391

[15] R. J. Baddeley, B. W. Tatler, High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis, Vision research 46 (18) (2006) 2824–2833.
URL http://www.sciencedirect.com/science/article/pii/S0042698906001192

[16] P. Reinagel, A. M. Zador, Natural scene statistics at the centre of gaze, Network: Computation in Neural Systems 10 (4) (1999) 341–350.
URL http://informahealthcare.com/doi/abs/10.1088/0954-898X_10_4_304

[17] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention, Vision research 42 (1) (2002) 107–124.
URL http://homepage.psy.utexas.edu/Homepage/Class/Psy394U/Hayhoe/perceptionaction/2008/week4readings/parkhurst.pdf

[18] B. W. Tatler, R. J. Baddeley, I. D. Gilchrist, Visual correlates of fixation selection: Effects of scale and time, Vision research 45 (5) (2005) 643–659.
URL http://www.activevisionlab.org/resources/papers/Tatler_etal_VisRes_05.pdf

[19] N. D. B. Bruce, D. P. Loach, J. K. Tsotsos, Visual correlates of fixation selection: a look at the spatial frequency domain, in: Image Processing, 2007. ICIP 2007. IEEE International Conference on, Vol. 3, 2007, p. III–289.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4379303

[20] A. A\ccık, S. Onat, F. Schumann, W. Einhäuser, P. König, Effects of luminance contrast and its modifications on fixation behavior during free viewing of images from different categories, Vision research 49 (12) (2009) 1541.

URL `http://www.sonat.bonobostree.com/doc/pdf/acik_vr_2009.pdf`

[21] D. Gao, N. Vasconcelos, Discriminant interest points are stable, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 2007, p. 1–6.
URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4270146`

[22] D. Gao, V. Mahadevan, N. Vasconcelos, On the plausibility of the discriminant center-surround hypothesis for visual saliency, Journal of vision 8 (7).
URL `http://ww.journalofvision.org/content/8/7/13.short`

[23] D. Zoran, Y. Weiss, Scale invariance and noise in natural images, in: 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 2209 –2216. `doi:10.1109/ICCV.2009.5459476`.

[24] Y. Weiss, W. Freeman, What makes a good model of natural images?, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07, 2007, pp. 1 –8. `doi:10.1109/CVPR.2007.383092`.

[25] G. A. Alvarez, A. Oliva, Spatial ensemble statistics are efficient codes that can be represented with reduced attention, Proceedings of the National Academy of Sciences 106 (18) (2009) 7345–7350. `doi:10.1073/pnas.0808981106`.
URL `http://www.pnas.org/content/106/18/7345`

[26] M. N. Do, M. Vetterli, Wavelet-based texture retrieval using generalized gaussian density and Kullback-Leibler distance, Image Processing, IEEE Transactions on 11 (2) (2002) 146–158.
URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=982822`

[27] M. N. Do, M. Vetterli, Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden markov models, Multimedia, IEEE Transactions on 4 (4) (2002) 517–527.
URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1176950`

[28] N. Minh, Directional multiresolution image representations, Ph.D. thesis, Citeseer (2002).
URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.162.3137&rep=rep1&type=pdf

[29] K. X. Shi, Mapping natural image patches by explicit and implicit manifolds, ProQuest, 2009.
URL http://books.google.co.uk/books?hl=en&lr=&id=zi6Ix_eUy9wC&oi=fnd&pg=PA1&dq=mapping+natural+image+patches+by+explicit+and+implicit+manifolds&ots=A4afgJaFDQ&sig=FXdjzAn_w65WjBAoAOTC2AgpAfI

[30] M. Crouse, R. Baraniuk, Simplified wavelet-domain hidden markov models using contexts, in: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998, Vol. 4, 1998, pp. 2277 –2280 vol.4. doi:10.1109/ICASSP.1998.681603.

[31] J. Romberg, H. Choi, R. Baraniuk, Bayesian tree-structured image modeling using wavelet-domain hidden markov models, IEEE Transactions on Image Processing 10 (7) (2001) 1056 –1068. doi:10.1109/83.931100.

[32] E. Simoncelli, J. Portilla, Texture characterization via joint statistics of wavelet coefficient magnitudes, in: 1998 International Conference on Image Processing, 1998. ICIP 98. Proceedings, Vol. 1, 1998, pp. 62 –66 vol.1. doi:10.1109/ICIP.1998.723417.

[33] J. Pesquet, H. Krim, D. Leporini, E. Hamman, Bayesian approach to best basis selection, in: , 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, Vol. 5, 1996, pp. 2634 –2637 vol. 5. doi:10.1109/ICASSP.1996.548005.

[34] F. Abramovich, T. Sapatinas, B. W. Silverman, Wavelet thresholding via a bayesian approach, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60 (4) (1998) 725–749. doi:10.1111/1467-9868.00151.
URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00151/abstract

[35] H. A. Chipman, E. D. Kolaczyk, R. E. McCulloch, Adaptive bayesian wavelet shrinkage, Journal of the American Statistical Association 92 (440) (1997) 1413–1421. `doi:10.1080/01621459.1997.10473662.` URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10473662`

[36] H. Cheng, C. A. Bouman, J. P. Allebach, Multiscale document segmentation 1, in: IS AND T ANNUAL CONFERENCE, 1997, p. 417–425. URL `https://engineering.purdue.edu/~bouman/publications/pdf/ist97.pdf`

[37] H. Cheng, C. Bouman, Trainable context model for multiscale segmentation, in: 1998 International Conference on Image Processing, 1998. ICIP 98. Proceedings, Vol. 1, 1998, pp. 610 –614 vol.1. `doi:10.1109/ICIP.1998.723575.`

[38] J. Li, R. Gray, R. Olshen, Multiresolution image classification by hierarchical modeling with two-dimensional hidden markov models, IEEE Transactions on Information Theory 46 (5) (2000) 1826 –1841. `doi:10.1109/18.857794.`

[39] C. Bouman, M. Shapiro, A multiscale random field model for bayesian image segmentation, IEEE Transactions on Image Processing 3 (2) (1994) 162 –177. `doi:10.1109/83.277898.`

[40] H. Choi, R. Baraniuk, Multiscale image segmentation using wavelet-domain hidden markov models, IEEE Transactions on Image Processing 10 (9) (2001) 1309 –1321. `doi:10.1109/83.941855.`

[41] H. Choi, R. Baraniuk, Multiscale texture segmentation using wavelet-domain hidden markov models, in: Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems amp; Computers, 1998, Vol. 2, 1998, pp. 1692 –1697 vol.2. `doi:10.1109/ACSSC.1998.751614.`

[42] P. Moulin, J. Liu, Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors, IEEE Transactions on Information Theory 45 (3) (1999) 909 –919. `doi:10.1109/18.761332.`

[43] A. Borji, D. N. Sihite, L. Itti, Quantitative analysis of Human-Model agreement in visual saliency modeling: A comparative study, Image

Processing, IEEE Transactions on.

URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6253254